Inferring Human History from the Joint Site-Frequency Spectrum

Ryan N. Gutenkunst†, Carlos D. Bustamante, Scott H. Williamson

Department of Biological Statistics and Computational Biology – Cornell University – †rng7@cornell.edu

Abstract

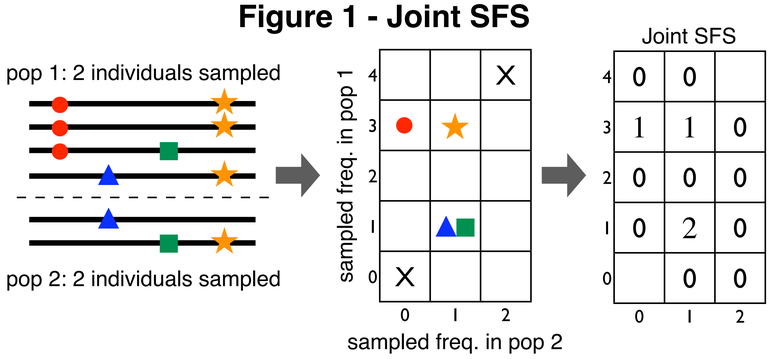
Our individual genomes are shaped by our ancestry. Similarly, contemporary genetic variation within and between populations is shaped by the history of those populations. We develop a novel statistical approach for inferring the history of interacting populations, based on the joint site-frequency spectrum of genetic variation. The efficiency of our approach allows us to consider complex histories of interacting populations in a statistically rigorous manner. We apply our method to human expansion out of Africa and into Europe and Asia, constraining our model using public data from the Environment Genome Project. Preliminary results suggest the existence of a distinct Eurasian population that mixed substantially with African populations before the divergence of Europeans and Asians.

Joint Site-Frequency Spectrum

Much of human genetic variation is comprised of single nucleotide polymorphisms (SNPs), locations in the genome where some individuals have one DNA base and others have another. We can compare with the chimp reference sequence to determine which possibility is ancestral and which is the derived mutation.

The distribution of SNPs between individuals and between populations can be summarized in the joint site-frequency spectrum (SFS). Given samples from N populations, the joint SFS is an N–dimensional matrix whose entries count the number of SNPs at a given frequency in each of the population samples. In the approximation that the SNPs are independent, this is a complete summary of the data. Construction of the joint SFS is illustrated in **Figure 1**, where SNPs are represented by colored shapes.

Figure 1 - Joint SFS

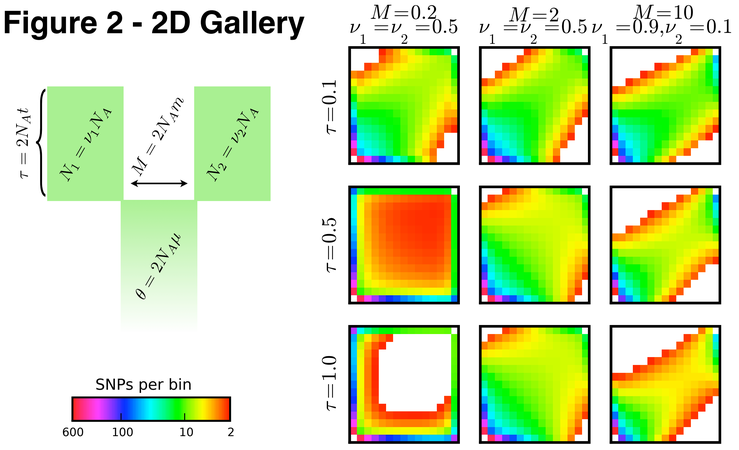


The joint SFS encodes information about the history of the populations. For example, if the populations diverged recently or have substantial migration between them, then SNP frequencies in the two populations will be highly correlated and the SFS will be diagonally dominated.

Joint SFS Gallery

Figure 2 shows the joint SFS between two diverging populations for several scenarios. The ancestral population size is N_A, and the relative sizes of populations 1 and 2 are ν_1 and ν_2, respectively. m is the fraction of individuals migrating each generation, while M and τ are the scaled migration rate and divergence time. Different histories can leave striking signatures in the joint SFS.

Figure 2 - 2D Gallery



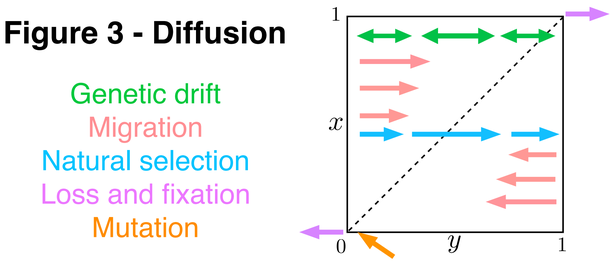
Simulation via Diffusion

To make inferences, we simulate frequency spectra using a diffusion approach and fit to data. For two interacting populations, we solve for the density, $\phi(x, y, t)$, at time t of SNPs at relative frequency x in population 1 and y in population 2. The evolution equation for $\phi(x, y, t)$ is:

$$\frac{\partial \phi}{\partial t} = \frac{1}{2}\frac{\partial^2}{\partial^2 x}\left[\frac{x(1-x)}{2N_x}\phi\right] - \frac{\partial}{\partial x}\left[\left(m_{x\leftarrow y}(y-x) + s_x\, x(1-x)\right)\phi\right] + \frac{1}{2}\frac{\partial^2}{\partial^2 y}\left[\frac{y(1-y)}{2N_y}\phi\right] - \frac{\partial}{\partial y}\left[\left(m_{y\leftarrow x}(x-y) + s_y\, y(1-y)\right)\phi\right]$$

The highlighted terms are illustrated in **Figure 3**. The green term models random frequency drift due to Wright-Fisher mating, while the pink term models migration, and the blue term models natural selection. Our boundary conditions include loss and fixation of mutation at the "corners" of the domain and influx of new mutations at low frequency in each population. Generalizing to more interacting populations is straightforward.

Figure 3 - Diffusion



Human Expansion Out of Africa

We apply our method to the expansion of humans out of Africa into Asia and Europe. We fit public resequencing data from the NIEHS Environmental Genome Project, focusing on Yoruba individuals from Ibadan, Nigeria, Han Chinese from Beijing, and Americans of European descent. The upper left panel of **Figure 4** shows the resulting 3D joint SFS. The top row of the right hand panels shows the 2D marginal spectra. Note that the Yoruba/European and Yoruba/Chinese spectra are very similar, suggesting that Europeans and Asians recently diverged from an earlier population that interacted strongly with Africa.

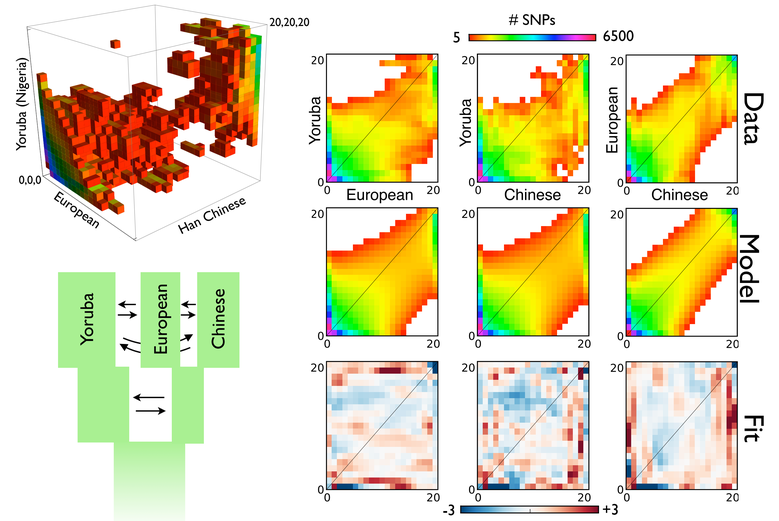


Figure 4 - Human data, model, and fit

To quantify this observation, we fit a divergence model with migration rate and population size changes at the divergence times. Our best fit model places the African/Eurasian split at about 130 thousand years ago, with significant migration until the European/Asian split about 17 thousand years ago. The bottom two rows of **Figure 4** show the 2D projection of the best-fit model and the residuals between the data and model. The correlated residuals indicate that our model is incomplete. In particular, there is strong evidence for recent European and Asian population growth.

To rigorously evaluate the goodness-of-fit of our model and the statistical significance of more complicated models, we fit complementary simulated data that accounts for the linkage between SNPs in our real data. This work is ongoing.

Acknowledgments

For assistance with the sequence data, we thank Ryan Hernandez. We also thank him and Adam Auton for many helpful discussions. The work was funded by Cornell University.